ECE 6885 - Reinforcement Learning Instructor: Prof. Tian Lan

HW #4 (Due: May 01, 2023)

Problem 1.

Consider an MDP with two states $\{A, B\}$ and two actions $\{a^1, a^2\}$. The system state transitions are governed through the following transition matrices:

$$M(a^1) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad M(a^2) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The reward is as follows:

$$\begin{cases} +5 & \text{moving to State } B \\ 0 & \text{moving to State } A \\ -1 & \text{taking action } a^2 \\ 0 & \text{taking action } a^1 \end{cases}$$

Consider an initial policy $\pi^0 = \begin{bmatrix} \pi^0(A) \\ \pi^0(B) \end{bmatrix} = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix}$, $\gamma = 0.9$ and episode length 5. Perform Monte Carlo Policy Iteration method to obtain the best policy.

* you need to show all trajectories, the approximation of Q-values and Policy Improvement till the time that Policies in two Consecutive Iterations stays the same.

Problem 2.

Consider the following system with two states $\{A, B\}$ and two actions $\{a^1, a^2\}$. The system state transition is unknown and learning should be achieved through interactions. Consider the following state-action-reward obtained through Softmax policy in Actor-Critic algorithm.

$(S_0 = A, a_0 = a^1, r = 10), (S_1 = A, a_1 = a^2, r = -5), (S_2 = B, a_2 = a^1, r = 40),$

$(S_3 = A, a_3 = a^2, r = -5), (S_4 = B, a_4 = a^2, r = 20), (S_4 = A, a_4 = a^1, r = +10), S_5 = A$

Set the initial preferences and state values to zero. Use $\alpha = 0.5$, $\beta = 0.1$ and $\gamma = 0.9$ and show all intermediate preferences, state values and policies.